Webaccess to the CESAR Observatory data: the CESAR DATABASE SYSTEM portal

Henk Klein Baltink[1], Berend Wilkens[2], Fred Bosveld[1], Mark Savenije[1]

[1]Royal Netherlands Meteorological Institute, Regional Climate Division, P.O. Box 201, 3730 AE De Bilt, The Netherlands
henk.klein.baltink@knmi.nl,fred.bosveld@knmi.nl,mark.savenije@knmi.nl
[2] Royal Netherlands Meteorological Institute, Information and Observation Services and Technology department, berend.wilkens@knmi.nl

ABSTRACT

The Cabauw Experimental Site for Atmospheric Research (CESAR) was established in 2002 as a consortium of eight Dutch institutes involved in atmospheric research [1]. The CESAR site is an atmospheric reference site of major importance in the European and medial atmospheric research networks. Observations at CESAR cover a wide spectrum of measurements and data are acquired by very different instruments on site and in its near surroundings. Instruments deployed include remote sensing systems, tower and surface instruments, and soil, hydrological and aerosol instruments. In the near future also infra sound will be measured at the surface and in the tower. The data collected by these systems is very different both in nature and content. Data is delivered to the user community both in near real time and offline.

To utilize the CESAR data to its full potential it is important for the data users, both for CESAR partners and other users, to be able to access, exchange and retrieve the data in an user friendly way with an acceptable response time and without any operator interference. A web based data management and distribution system is crucial to obtain this goal. Therefore KNMI (Royal Netherlands Meteorological Institute) developed during the last two years in-house the CESAR database system (CDS) to address the need for CESAR data delivery and distribution.

The frontend (web portal) of the CDS comprises a self-registration, a search function with a limited number of search options, and a download basket for ordering data files. Metadata on the dataset level is provided in the web portal (see section 2.), as well as an option to preview the data by the quicklooks which are stored with the data files in the database. The data itself are all stored in netCDF format [2] and comply with the Climate and Forecast Metadata Convention version 1.4 [3]. The backend of the system includes ftp-buffers for upload of the data by the data providers and for near real time distribution also by ftp. The core of the backend is based on the Netherlands Atmospheric Data Centre (NADC) software [4], which handles all the ingestion, archiving, metadata extraction, and the generation and distribution of the logging and error reports. An offline facility has been implemented to retrieve statistics about user visits and download orders.

After some months of testing the CDS has become operational in July 2009 and will in the course of the next half to one year be filled with many datasets from the operational continuous measurement program, measurements executed by visiting researchers and datasets from (inter)national campaigns hosted at CESAR.

In section 1 we describe the general set-up of the CDS in more detail. In section 2 the metadata both at the level of the dataset and datafile are discussed. Section 3 presents an overview of the main functionalities of the CDS web portal.

1. CDS GENERAL DESCRIPTION

The first initiative to build a CESAR database dates from as far back as 2003 when a user survey was conducted amongst the CESAR data providers. The results of this survey formed the starting point of the development of the CDS. Amongst others because of limited resources and constraints concerning the complexity of the database system a solution of medium complexity was designed. Open source software and existing applications at KNMI were used where feasible for the realisation of the CDS.

The database is not developed as a full scale relational database, but rather provides the user a search facility with a limited number of parameters. The basic layout of the CDS is schematically presented in fig. 1.

The data files and associated quicklooks are stored in a filesystem (NADC Archive). The metadata of the data files are stored in a MySQL database (CDS Metadata). The authentication and authorisation is implemented using the Apache Tomcat webserver generic access authorisation tool. In principle the datasets available in the CDS are free for non-commercial use under the restrictions set in the CESAR data policy document. However it is possible for the manager of the CDS to restrict access to a limited user group. Restricted access can be set at the level of the dataset only. All datasets, the public and the restricted, are delivered free of any charge. Therefore no accounting is implemented, only for creating user and download statistics a log of all access and downloads is maintained.

The hardware of the CDS is embedded in the ICT infrastructure of the KNMI. The CDS runs on a cluster of two blade servers to ensure high availability. The input and output buffers are located on the main ftp-server of the KNMI, the harddisks for the storage of the data files and MySQL databases are part of KNMI's Central Data Storage system connected to the webserver via high speed optical fibre links. This set-up allows for easy extension of the harddisk capacity, and furthermore the system status monitoring is now fully incorporated in existing monitoring services. Again to ensure a high and long-term availability of the CDS.

© Proceedings of the 8th International Symposium on Tropospheric Profiling, ISBN 978-90-6960-233-2
Delft, The Netherlands, October 2009. Editors, A. Apituley, H.W.J. Russchenberg, W.A.A. Monna

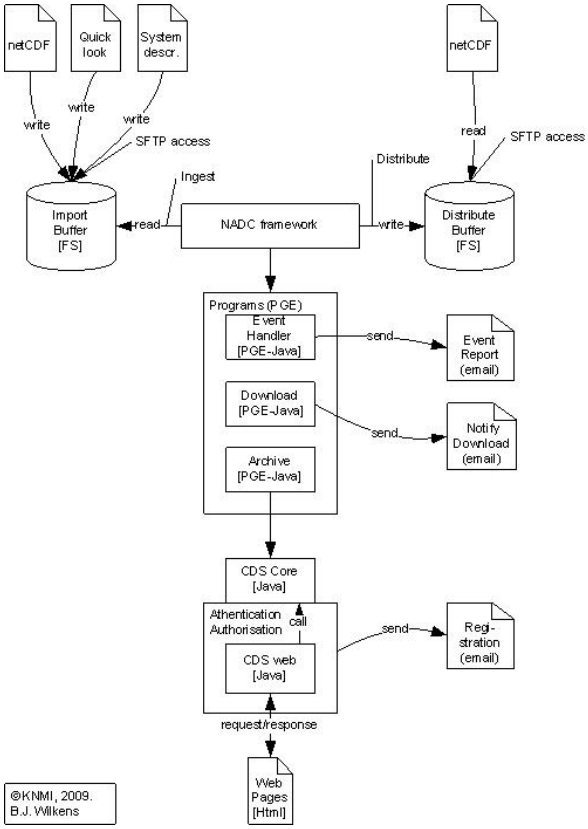

Figure 1. Layout of the main components of the CDS.

Software developed for the CDS is based on open source domain packages. An overview of the different sources used is presented in figure 2. To generate reports on the user logon and download statistics a web based utility was developed in PHP. In a later stage this offline utility will be integrated in the CDS.

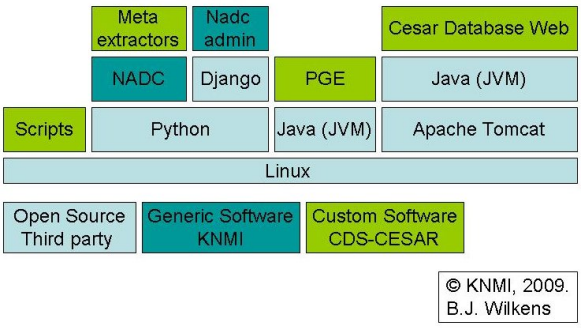

Figure 2. Software packages used for the CDS

The main processing unit, the NADC framework is a general purpose software package developed at KNMI for processing, archiving and managing satellite data, but is also easily applicable to other data stream management applications like the CDS. The NADC system is build by KNMI and SRON (Netherlands Institute for Space Research) and has its origin from previous projects like the Netherlands Sciamachy Data Center, OMI Dutch Processing System and GOME-2 Processing System. The NADC system is developed using community software and open standards where possible like MySQL, Python and OpenSSH. The NADC handles all the main tasks like file ingestion, file distribution, message generation, scheduling, etc. Only the datastream specific modules i.e. the metadata extractors, scripts, website and PGEs (Product Generation Executives) needed to be developed specific for the CDS.

2. DATASET AND DATAFILE METADATA

In data intensive environments like atmospheric research, it is mandatory to describe data in an unambiguous manner to:

1) find the data (*discovery*);

2) assess the quality of the data (*exploration*);

3) use the data in a correct manner (*exploitation*)

Metadata are data describing the data and are essential for the users either to find the required data and/or to assess the usefulness of the data (quality, origin, etc). Metadata can apply to different levels of data, e.g. the whole dataset or to a single member of the dataset. Common terms are 1) *catalogue:* the collection of all categories; 2) *category:* a collection based on a certain theme or subject, e.g. water temperature, ozon, surface observation, a category contains datasets or other categories; 3) *dataset:* a collection of one product e.g. all files from the Cabauw tower dataset, and 4) *product:* smallest possible granule of a dataset, e.g. Cabauw datafile for one day.

The international accepted standard for metadata for geospatial data is ISO 19115:2003 which includes approximately 400 metadata-elements. This is a content standard which determines which elements are described, but not how these needs to be implemented technically. For Dutch governmental organizations a Dutch profile has been created based on the ISO 19115:2003 set. This profile is maintained by Geonovum [5]. The use of this profile will become mandatory for Dutch governmental organizations who want to share data in an interoperable manner. This profile applies to the level of the dataset, but not to the products in the dataset.

The metadata on the dataset level are available in the web portal for each published dataset. The metadata elements are based on the WMO core set. Recent projects like ADAGUC [6] and INSPIRE [7] have defined also metadata lists to be used for geospatial datasets. For the metadata at the file level we have chosen to comply with the definition and standards used in the NetCDF Climate and Forecast Metadata Convention [3], a widely used convention in the atmospheric community.

All the datafiles stored in the filesystem of the CDS must comply with the CDS standard filename. The filename format allows the user to easy recognize the contents of the file from the filename itself. CDS files comply with the following filename convention:

<source>_<system>_<parameter>_<level>_

<repeat-period>_<version>_<dtg>.<ext>

where:

source: a name which describes the origin of the file, here mostly 'cesar'; *system_parameter:* unambiguous name or abbreviation representing the data file content, eg. tower_fluxes; *level:* indicates the level of processing applied to the data; *repeat-period:* indi-

cates the typical cycle interval (or averaging interval) of the data (default in minutes, but seconds and hours can be used too by adding an 's' or 'h' resp.); *version:* version number of data; and *dtg:* a date-time identifier for the start of data interval in the format 'yyyy[mm[dd]]'. An example for a CT75 ceilometer data file: "*cesar_ct75ceilometer_back-scatter _la1_t30s_v1.0_20090812.nc".*

The time variable in the file is always in hours offset from midnight (UTC) from the day that follows from the filename. E.g. if the file has *_200901.nc as filename, the time variable is in hours referenced to 20090101 midnight UTC. This is stricter than the CF convention prescribes. In addition to the 'standard_name' attribute of the CF convention, an additional 'cesar_standard_name' attribute can be defined for commonly used variables in the CDS datasets.

3. THE CDS WEBPORTAL

3.1 Self-registration

The CDS is developed to access the data without any operator interaction. In principle datasets are accessible to all users who agree with the CESAR data-policy for data use. A straightforward self-registration procedure is available in the web portal. On first sign-on users fill in a valid email-address, their full name and affiliation [8]. To prevent automated registration by bots the user needs to enter a 'captcha' code [9]. After successful registration a password is sent to the user's email address. On self-registration the user gets access to all public datasets. Access is regulated by a role; each user is assigned initially a public role which gives access to all public datasets. Access to restricted datasets can only be granted by the manager of the CDS by giving the user a different role. For example if the radar data are restricted the user will be assigned a role that both gives permission to access all public data and the radar data. A role can have multiple permissions, but each dataset is linked to only one permission. Once a user is registrated the user can view his user account detail in the "My Cesar" window.

3.2 Browsing datasets.

The CDS provides two methods for browsing the datasets, i.e. 1) using the search facility for combined keyword, time and characteristic value queries, or 2) using the category tree. The latter provides a graphical representation of the CDS by category in the form of a collapsible tree view.

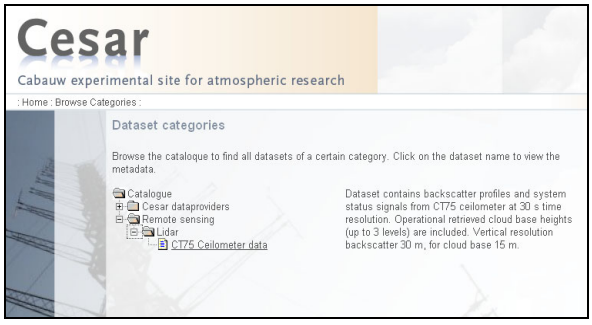

Figure 3. The collapsible category tree view, on mouse over on an item some metadata is displayed.

In the search facility the 'unit' of the search query is a single day. For each day up to 20 daily characteristics values can be provided for search e.g. mean air temperature, or daily accumulated precipitation. The 20 parameters are selected by the manager of the CDS after consulting the data-providers. The keywords are based on a WMO keyword list, but can be extended with any word deemed necessary by the data provider to characterise their dataset.

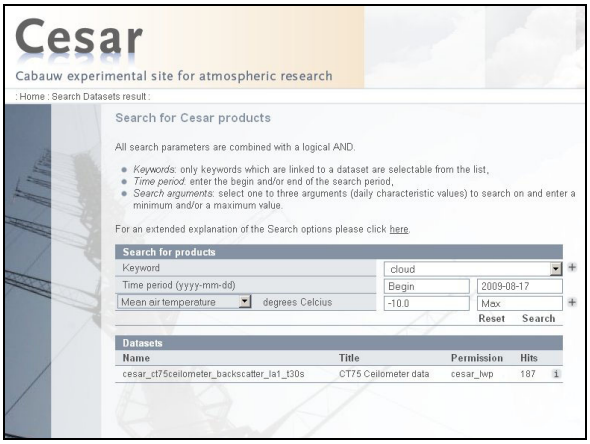

Figure 4. The CDS search window

Once a selection has been made the user can browse through the selection and preview the quicklooks of each data product. In the preview window there is an option to add the data product to the download basket. The search selection can also be modified by changing the parameters in the search option window. The values of the last search are kept in memory and don't need to be entered again. In order to make the CDS system not too complex it was decided not to implement 'on-the-fly' creation of quicklooks and use fixed sized prepared quicklooks only. One quicklook per data product can be shown, an example is shown in fig. 5.

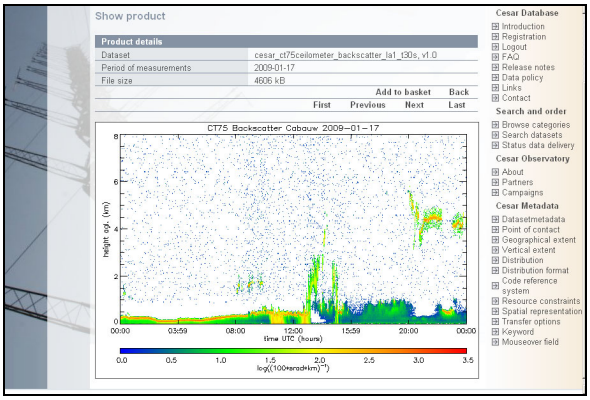

Figure 5. Example of quicklook preview of a data product.

3.3 Ordering data products

If a selection is made and the user wants to order the selected data products he can add the selection to the download basket. As long as the basket is open, i.e. is not ordered, products can be added or deleted. The

total size of the data products in the basket is limited to 5 GB uncompressed. Once a basket has been ordered the CDS will start a process of retrieving the files from the archive, compress the files using gzip, and make a single tarfile containing all the products requested in the order. Upon completion of this task the CDS will send an email to the user with a link to the "My Cesar" page of the user on which a clickable link to the order is available. The order will be available for download for a week, after this week the tar file is deleted from the disk. The status of the recent orders can be viewed in the "My Cesar" window (fig. 6)

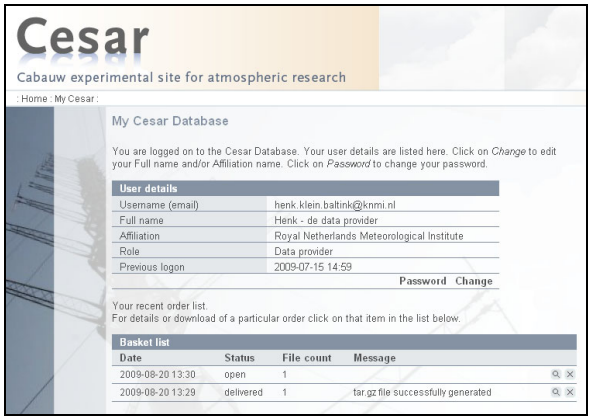

Figure 6. The "My Cesar" window with the status of recent download orders.

4. SUMMARY

To provide access to the CESAR data a versatile CESAR database system with web portal (CDS) has been developed. Self-registration to the CDS provides the user access to all public datasets without any operator interaction. A relative simple search facility is available to make a selection from the datasets stored. All the datasets have associated metadata to describe their contents. The datafiles are netCDF formatted and the content of the file complies with the CF1.4 convention on metadata. Data upload by CESAR data providers and distribution at the backend of the CDS is fully automated.

REFERENCES

[1] CESAR: http://www.cesar-observatory.nl.

[2] netCDF - network Common Data Form: http://www.unidata.ucar.edu/software/netcdf

[3] CF - NetCDF Climate and Forecast Metadata Convention: http://cf-pcmdi.llnl.gov

[4] NADC: http://neonet.knmi.nl/neoaf

[5] Nederlands profiel op ISO 19115 voor geografie, v1.2, (http://www.geonovum.nl)

[6] ADAGUC: http://adaguc.knmi.nl

[7] INSPIRE: http://www.inspire-geoportal.eu

[8] CDS: http://www.cesar-database.nl

[9] captcha: http://www.captcha.net